EX/PG/CSE / T / 128F/8/2015 Text Analytics

Master in Computer Science & Engineering, Second Semester, Paper VIII

Overview: This subject introduces the theory and methods of text processing applications, such as information retrieval, question-answering, automatic summarization, and evaluation of the desired systems. The course covers knowledge-based and statistical approaches to language and text processing for post-graduate students.

Course Outcomes - Upon successful completion of this course, students will be able to demonstrate accomplishments of knowledge and comprehension, application and analysis, and synthesis and evaluation:

Module 1: Information Retrieval and Extraction [20L]

Boolean retrieval; Term vocabulary and postings lists; Dictionaries and tolerant retrieval; [2L]
Index construction; Index compression; Scoring, term weighting and the vector space model; Computing scores in a complete search system; [4L]
Advanced topics in Information Retrieval: Relevance feedback and query expansion; XML retrieval; Probabilistic information retrieval; Language models for information retrieval. [3L]
Machine Learning in Information Retrievall: Text classification and Naive Bayes; Vector space classification; Support vector machines and machine learning on documents. [4L]
Association and clustering: Apriori, **K-Means, FCM**, Hierarchical; Matrix decompositions and latent semantic indexing. [5L]
Web Mining: Web search basics; Web crawling and indexes; Link analysis. [2L]

Module 2: Question Answering [8L]

Open and Restricted Domain Question-Answering; Question Classes and Processing; Context and Data Source; [3L]
Extraction and Formulation of Answers; Web exploitation: Real Time, Multilingual and Interactive QAs; [2L]
Textual Entailment: Learning to recognize features; Textual Inference and Semantic Entailment; [3L]

Module 3: Summarization [8L]

Automatic Summarization: Factors and Directions, Extraction, Abstraction, Maximum Entropy and Aided Summarization; [3L]
Keyphrase Extraction; Supervised and Unsupervised Learning for Sentence Compression, Sentence Fusion [2L]
Document Summarization: Single and Multidocument; Web page and form summarization, clickthrough data, Summarization on Online Discussions and Blogs. [3L]

Module 4: Evaluation [6L]

Precision, Recall and F-Measure, Confusion Matrix; MAP, MRR [2L]
Quantitative and Qualitative Evaluation, Accuracy and Question Inversion, Answer Selection, Assessment of Reliability and Validity, Open vs. Restricted Domain; [2L]
N-gram Co-Occurrence Statistics for Summarization; Measures : Intrinsic (Coherence, Informativeness, Utility Method, Content Similarity, BLEU and ROUGE), Extrinsic (Keyword Association and Gaming), Inter and Intra Textual Evaluation. [2L]

Reference*:* Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Introduction to Information Retrieval, Cambridge University Press. 2008.